DOCUMENT RESUME

ED 137 337 TM 006 138

AUTHOR Rudner, Lawrence M.
TITLE An Approach to Biased Item Identification Using Latent Trait Measurement Theory.
PUB DATE [Apr 77]
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Bias; Criteria; Culture Free Tests; *Item Analysis; Item Banks; *Mathematical Models; Probability; *Test Bias; True Scores
IDENTIFIERS *Item Characteristic Curve Theory; Latent Trait Theory

ABSTRACT

Because it is a true score model employing item parameters which are independent of the examined sample, item characteristic curve theory (ICC) offers several advantages over classical measurement theory. In this paper an approach to biased item identification using ICC theory is described and applied. The ICC theory approach is attractive in that it, (1) appears to be sensitive largely to cultural variations in the trait gauged by test items, (2) does not assume total scores to be valid indicators of true ability, (3) places the identified degree of item bias on a quantified metric, and (4) is applicable to items of sufficiently varying degrees of difficulty. While sensitive to some factors other than item bias, namely, local independence, item inappropriateness and poor parameter estimates, the approach may prove useful to the measurement field. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED137337

AN APPROACH TO BIASED ITEM IDENTIFICATION USING LATENT TRAIT MEASUREMENT THEORY

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Lawrence M. Rudner
Gallaudet College
Model Secondary School For The Deaf
Washington, D.C. 20002

Paper Presented At The Annual Meeting Of The American Educational Research Association, New York, April, 1977

TM006 138

Over the past several years, the issue of bias in intelligence testing, achievement testing, and testing for selection and placement has been of increasing concern to both the layperson and the measurement expert. In response to this concern, various models have been proposed for evaluating bias both in a measure as a whole and in the items within a measure. Models for evaluating bias in measure as a whole (see the Spring 1976 issue of the Journal of Educational Measurement) are of primary interest to the test user as they assist in the fair use of test results. Models for evaluating the items within a measure (see the reviews by Merz, 1976 and by Rudner, 1977a) are of prime interest to the test developer. These approaches have the potential to assist in developing valid and cross-culturally fair test items. This paper is addressed to an improved method for analyzing item bias.

SOME CRITERIA FOR AN IMPROVED APPROACH

In their reviews of the literature, Merz and Rudner discussed several of the approaches to biased item identification along a variety of dimensions. Although the intent of these discussions was to identify relative merits and weaknesses, some of the dimensions can be used to establish criteria for an improved approach. The following criteria and rationales are proposed.

The author is indebted to David Knight for his valued input on earlier drafts of this report and to the Office of Demographic Studies at Gallaudet College and an anonymous West Coast school district for providing the data used in the study.

An improved approach to biased item identification should:

1. be sensitive only to group differences in the factor gauged by the item

   Item bias is concerned with whether an item measures the same trait across populations. An improved approach should identify only items which fail to do this and not be overly sensitive to factors other than bias, e.g. group differences in ability, sampling, and item inappropriateness.

2. not assume total scores to be valid indicators of ability

   Total observed scores are obtained by summing item responses. Consequently, the presence of biased items causes one to suspect that the total scores contain additional error. An approach relying on this assumption could yield spurious results.

3. quantify degree of item bias

   While it is convenient to refer to an item as being biased or unbiased, this dichotomous distinction can be infexible as well as misleading. The investigator needs to be able to vary the definition of what is "very biased" to suit the purposes of the study. An improved approach must at least have this flexibility and preferably map the degree of item bias to a meaningful scale.

4. be applicable to items of varying difficulty

   Some of the previously proposed approaches are limited in their ability to detect item bias in easy or difficult items. While no approach can be expected to detect item bias when almost all or none of the examinees respond correctly, an improved approach should, at least, be applicable over a wide range of item p-values.

This paper describes an approach which capitalizes on item characteristic curve (icc) theory and employs a definition of bias similiar to that used by Green and Draper (1972), Scheuneman (1976), and Pine and Weiss (1976). This icc theory approach appears attractive when measured against the above criteria.

A BRIEF OVERVIEW OF ICC THEORY

Latent trait or item characteristic curve theory relates the probability of a correct item response to a function of an examinee's underlying ability level (θ_i) and characteristic (s) of the item. While the various models (Lord, 1952; Rasch, 1960; Birnbaum, 1968; Urry, 1970) differ in terms of the number of item parameters considered; they all describe the item parameter (s) independently of the examined sample. This attractive property has lead to the development of some interesting applications in test development, adaptive testing and equating and may prove useful in detecting item bias.

One general, cumulative logistic model formalized by Birnbaum uses three item parameters: a_g - an item discrimination index, b_g - an item difficulty index, and c_g - a pseudo guessing parameter. Using the notation $P(u_g=1|\theta_i)$ to represent the probability of a correct response to item g given an examinee of ability level θ_i, Birnbaum's three parameter model states that:

$$P(u_g=1|\theta_i) = c_g + (1 - c_g)\ [1+\exp(\ -1.7a_g\ (\theta_i - b_g)\ )]^{-1}$$

This relationship between θ_i and $P(u_g=1|\theta_i)$ is illustrated in Figure 1.

---

insert Figure 1 about here

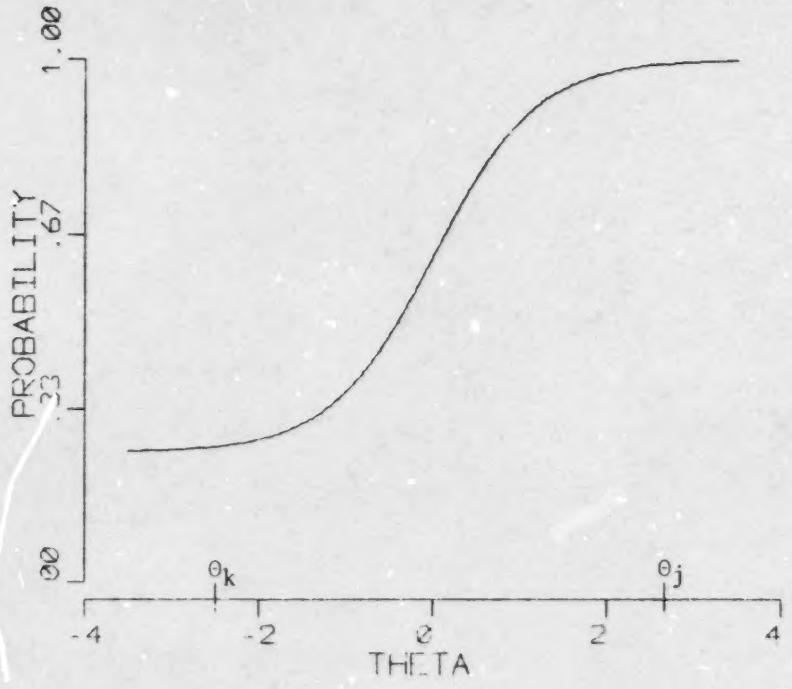

Figure 1: A hypothetical item characteristic curve

The probability of a correct response given a specific ability level increases monotonically as true ability increases. For example, an examinee with a high true ability, e.g. θ_j, has a high probability of responding correctly [$P(u_g=1|\theta_j)\rightarrow 1.0$]. Conversely, an examinee of low true ability, e.g. θ_k, has a low probability of responding correctly; approaching the lower asymptote of the curve, c_g.

The inflection point of the curve, b_g, is referred to as the item difficulty parameter in that it indicates the relative position of the curve along the θ axis. The more the curve is positioned to the right, the more ability is necessary for an examinee to have a good probability of a correct response. The slope of the curve at b_g helps define a third parameter, a_g. This value, referred to as the discrimination parameter, indicates the power of the item to separate examinees of close but unequal levels of ability. Although the item parameters and θ are on a common metric, these item parameters describe characteristics of the item independently of the examinee group. Full explanations and development of this and other mental measurement models can be found in Jensema (1972) and in Lord and Novick (1974).

ICC THEORY AND BIASED ITEM IDENTIFICATION

The only previous applications of icc theory for identifying biased items found in the literature were those of Green and Draper, Wright et.al. (1976) and Lord (in press). Green and Draper had used observed total scores as estimates of examinees' abilities, θ_i's, and the proportions of examinees responding correctly at each total score level as estimates of $P(u_g=1|\theta_i)$. Their procedure called for plotting estimated icc's for each item separately for each culture group and comparing the plots.

By this and other latent trait theory approaches, an item is unbiased if examinees of the same ability level, but of different cultural affiliations, have equal probabilities of responding correctly. That is, an item is unbiased if the estimated icc's obtained from the various culture groups are identical. As an example of a biased item, consider the two hypothetical curves shown in Figure 2. These curves are based on responses by two different culture groups to the same item. Total observed scores are used as estimates at θ_i and proportions of examinees responding correctly are used as estimates of $P(u_g=1|\theta_i)$. The curves are not identical, since the location parameters for the two curves are not equal. Such an item can be considered biased in that often examinees of the same ability level, e.g. $X_j = 58\%$, but from different culture groups, do not have similiar proportions of correct responses. [1]

---
insert Figure 2 about here

While this approach is appealing, it fails to meet the second and third criterion. The approach as used by Green and Draper directly incorporates total observed scores and quantification of the degree of item bias is difficult (an eyeballing procedure is used to identify a "very biased item").

---

1. In a recent Monte-Carlo investigation of test bias models, Pine and Weiss (1976) used a similiar operational definition of bias. Specifically they maintained equal Birnbaum a_g parameter values between groups and varied the b_g parameter values to vary the amount of bias.

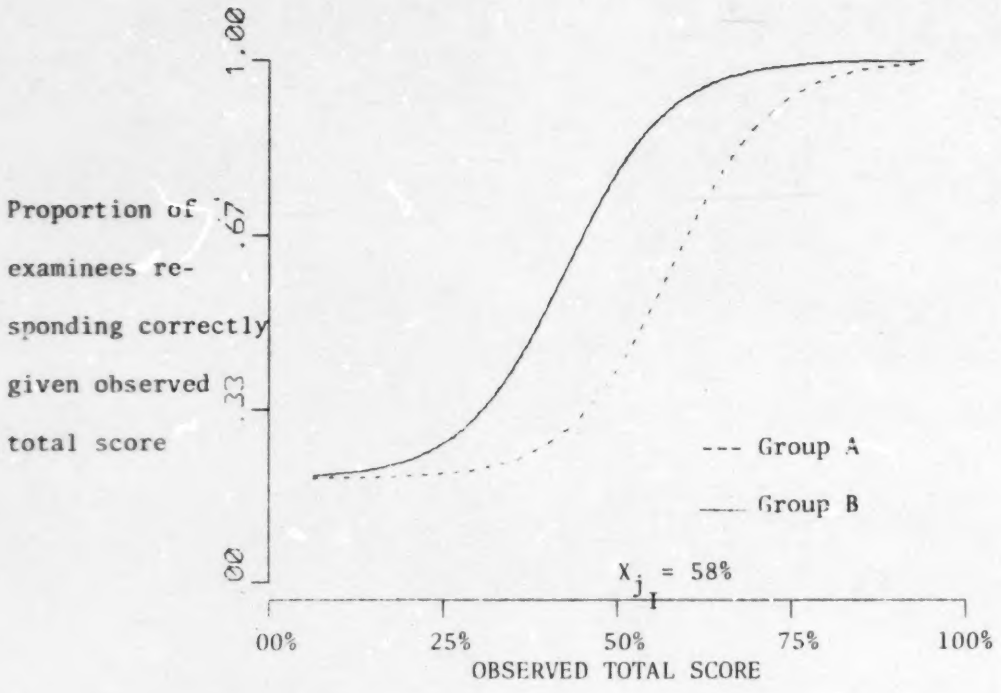

Figure 2: Two hypothetical response distributions

Rather than using total observed scores as estimates of θ_i and proportions as estimates for $P(u_g=1|\theta_i)$, more accurate values can be obtained using one of the recent methods of parameterization (Urry, 1975; Wingersky and Lord, 1973). During parameterization, the metric used for the θ scale is defined by the ability variance in the examined sample. In order to compare parameters obtained from two different examinee groups, the obtained values must be equated. Lord and Novick (1974, Chapter 16.11) and Rudner (1977b) have shown that this can be accomplished by computing the regressions of the parameter values based on one group of examinees on the parameter values based on the other group of examinees. The equated icc's will be identical when the restrictions of the model are met. That is, when the measure:

(1) is unidimensional

(2) contains locally independent items

(3) has error free parameter estimates

Since test items, by design, are usually locally independent and since accurate parameter estimates can usually be obtained, non-identical equated icc's would be largely indicative of non-unidimensionality. That is, aberrant equated icc's would indicate that the item measures (a) different traits across cultures (bias) or (b) a trait other than that gauged by the other items (inappropriateness).

One could evaluate the residuals from the regressions to gauge the extent of item bias. Rather than using residuals of the form $\hat{Y}-Y$, which cannot be readily compared between applications, perpendicular item - regression line distances could be used. Such a method would similiar to that suggested by Angoff (1972) for use with a regression-like technique whereby transformed p-values are examined for differences between groups.

An alternate approach is to refine the procedure used by Green and

Draper and use the equated parameter values to plot icc curves for each item by each culture group. The resultant scale would not dependent upon potentially biased total scores and $P(u_g=1|\theta_i)$ would be more accurately represented. However, eyeballing would still be necessary and the third criterion requiring quantification would not be met.

Wright, Mead and Draba (1976) have described an approach using the Rasch one parameter (b_g) model whereby goodness of fit residuals are examined for between - group differences. A more attractive approach was used by Lord (in press) who tested for a significant difference between equated icc's. An asymptotic significance test based on the summed variance - covariance matrices of the a_g and b_g parameter estimates was employed.

The approach preferred by author is to compute the area between the two equated icc's. This value would be low for relatively unbiased items and high for relatively biased ones. In most instances this area, defined by:

$$\phi_g = \int_{-\infty}^{\infty} [\,|\, P(u_g=1|\theta_i) - P'(u_g=1|\theta_i) \,|\,] \, d\theta$$

where $P(u_g=1|\theta_i)$ and $P'(u_g=1|\theta_i$ define the equated icc's for the two groups,

can be readily approximated on a high speed computer by:

$$\phi_g = \sum_{-5.000}^{5.000} [\,|\, P(u_g=1|\theta_i) - P'(u_g=1|\theta_i) \,|\,] \, \Delta\theta$$

where $\Delta\theta = .005$

This method places bias on a ratio scale and overcomes the problems of eyeballing and simultaneously analyzing differences in item discrimination and difficulty.

TWO APPLICATIONS OF THE ICC THEORY APPROACH

The icc theory approach to biased item identification is illustrated for two different situations. The first illustrates the approach when there are no biased items in the item pool. The second represents the approach as it might be used in test development. In the first situation, examinees from one culture group were randomly divided into two groups of different mean ability. Thus, two groups of the same cultural affiliation but different levels of ability were formed. Treating these groups as though they represented different cultures and applying the icc approach resulted in a pseudo-culture group comparison similiar to that employed by Jensen (1973) in evaluating an analysis of variance approach to biased item identification. Since both groups are of the same cultural affiliation, item aberrance should be minimal.

Item Pools - The 1973 Stanford Achievement Test, Form A, Primary 2 Battery, Reading Comprehension Subtest (SAT), -- which, item for item, is equivalent to the Stanford Achievement Test - Hearing Impaired Version, Level 2, Reading Comprehension Subtest -- formed the initial item pool for use in this study. The SAT consists of 16 paragraphs with a total of 48 four-choice items. According to the test publishers, emphasis is placed on comprehending disconnected discourse. It was anticipated that the SAT would contain several items biased in favor of one of the sampled culture groups.

Subjects - The study incorporates item responses made by large samples of examinees from two diverse culture groups. The first is composed of students in United States programs for the hearing impaired. The second is representative of the population for which the SAT was designed; namely normal hearing students in public schools. One major difference in these groups

is their exposure to and ability to use the English language (see Stokoe, 1976 for an excellent discussion of the social and cultural aspects of the deaf community).

In 1975, as part of the Annual Survey of Hearing Impaired Children and Youth, the Office of Demographic Studies at Gallaudet College collected item responses to the entire Stanford Achievement Test - Hearing Impaired Version. From their national random sample of 6,182 hearing impaired students, the sample of 2,637 examinees taking the Level 2 battery was extracted.

One thousand, six hundred three (1,603) students enrolled in a large West Coast public school district taking the SAT in the Spring of 1976 composed the sample of examinees representative of the population for which the measure was developed.

Procedures

The steps involved in applying the icc theory approach are:

1. Parameterize on each group separately (Urry's iterative minimum chi square technique was used)
2. Equate the scales by
   (a) regressing the a_g parameters obtained for the first group, through the origin, on the a_g parameters obtained for the second group, and
   (b) regressing the b_g parameters obtained for the first group on those obtained for the second[1]

[1]The magnitude of the R^2 inversely reflects the aggregate amount of aberrance. When the R^2 is low and hence many aberrant items are present, it is wise to trim items and recompute the regressions. This will prevent extremely biased items from overly distorting the regression equations used to equate the icc's.

3. The indicator of the degree of bias for each item g is the area between the equated icc's which is approximated by

$$\phi_g = \sum_{-5.000}^{5.000} [|P(u_g=1|\theta_i) - P'(u_g=1|\theta_i)|] \Delta\theta$$

where $\Delta\theta=.005$

For the pseudo group comparison, the hearing impaired examinees were randomly divided into two groups with different mean observed scores. This was accomplished by specifying, a priori, the desired observed score distributions of the two group of examinees. The resultant numbers of examinees were then converted to proportions of the total number of examinees needed for group assignment and proportions of examinees needed for each group. For each examinee, a random number was drawn and compared with the appropriate proportions to determine group assignment. A total of 528 examinees were lost due to the over abundance of examinees of certain observed score levels.

A PSEUDO-CULTURE GROUP COMPARISON

Summary statistics for the two pseudo-culture groups are shown in Table 1. The groups differed in mean observed scores thus implying differences in group ability.

---

insert Table 1 about here

Table 1

Test Statistics for the Two Pseudo-Culture Groups on the SAT

	N	$\bar{x}$	s.d.	KR-20
Group 1	1079	23.7	7.43	.83
Group 2	1030	20.9	6.97	.81

The equated and unequated parameter value estimates and the identified degrees of aberrance are shown in Table 2. For ease of interpretation the identified degrees of aberrance are plotted in Figure 3.

---

insert Table 2 and Figure 3 about here

The reader should note that with the exception of items 28 and 39 (and items 21 and 44 which had negative point biserials and could not be parameterized) all the identified degrees of aberrance are low, falling below .4. This value can be viewed as representing measurement noise in the form of parameterization error and slight deviations from unidimensionality and local independence.

A closer examination of the more aberrant items provides some added insight. Items 28 and 39 were more aberrant because of local dependence, non-within group unidimensionality or poor parameter estimates. The b_g parameters for these items were extremely high for the second group of examinees, namely 2.77 and 3.91 respectively. This can be loosely interpreted as meaning that, ignoring guessing, an examinee's ability must be 2.77 (3.91) standard deviations above the group mean ability to have a better than average chance of responding correctly. Since relatively few examinees were of this ability level, parameterization became tenuous and it is felt that the slight aberrance of these items was due to abnormally high parameterization error.

A DIVERSE CULTURE GROUP COMPARISON

Summary statistics for the two diverse culture groups are shown in Table 3. The equated and unequated parameter value estimates and the

Table 2

Equated and Unequated Parameter Estimates and
Degrees of Aberrance for the Pseudo-Culture Group Comparison

Item #	Group 1			Group 2 (Equated)			Group 2 (Unequated)			Aberrance
	a	b	c	a	b	c	a	b	c	
1	.82	- .22	.20	.88	- .07	.20	.93	.26	.27	.12
2	1.26	-1.42	.31	1.02	-1.25	.31	1.08	-1.10	.28	.15
3	.99	-1.32	.26	1.03	-1.18	.26	1.09	-1.02	.28	.10
4	1.24	.23	.38	1.24	.14	.38	1.31	.50	.35	.06
5	1.95	-1.54	.37	1.97	-1.41	.37	2.08	-1.28	.39	.08
6	.64	.83	.20	.68	.47	.20	.72	.88	.19	.28
7	1.02	.42	.18	1.21	.71	.18	1.28	1.16	.31	.24
8	.64	.93	.28	.64	.67	.28	.67	1.11	.25	.19
9	.79	.19	.11	.66	.30	.11	.70	.69	.11	.19
10	.96	1.40	.33	1.10	1.33	.33	1.16	1.87	.38	.08
11	.81	.10	.17	1.03	.16	.17	1.09	.52	.24	.18
12	1.02	- .12	.14	.91	.07	.14	.96	.42	.19	.17
13	1.42	-1.10	.29	1.38	-1.05	.29	1.45	- .87	.32	.04
14	1.09	.66	.33	.78	.55	.33	.82	.97	.24	.21
15	1.03	- .82	.21	1.00	- .77	.21	1.05	- .54	.24	.04
16	1.10	1.45	.32	.79	1.28	.32	.83	1.81	.25	.22
17	1.39	1.56	.33	.89	1.95	.33	.94	2.58	.31	.31
18	.91	1.78	.35	.63	1.60	.35	.66	2.18	.25	.26
19	.62	- .54	.13	.76	- .22	.13	.80	.09	.15	.32
20	1.01	1.84	.32	.76	2.14	.32	.80	2.80	.33	.24
22	1.49	3.68	.42	1.42	3.38	.42	1.50	4.23	.42	.17
23	.83	- .19	.28	.80	- .67	.28	.84	- .43	.20	.34
24	.75	- .60	.15	.94	- .57	.15	.99	- .31	.22	.19
25	.28	1.91	.30	.24	2.79	.30	.25	3.55	.30	.36
26	.68	.79	.26	.83	1.00	.26	.88	1.49	.36	.21
27	1.58	2.04	.27	1.77	1.89	.27	1.87	2.51	.28	.11
28	.70	3.01	.40	.72	2.11	.40	.76	2.77	.39	.51
29	1.02	.90	.28	.93	.76	.28	.98	1.21	.24	.11
30	1.11	2.03	.24	1.13	2.22	.24	1.19	2.90	.27	.14
31	1.16	.97	.27	1.09	1.09	.27	1.15	1.59	.31	.09
32	.94	.47	.16	.94	.55	.16	.99	.97	.25	.07
33	.89	1.85	.28	1.58	1.53	.28	1.67	2.10	.37	.34
34	1.80	- .25	.22	1.53	- .07	.22	1.61	.26	.29	.14
35	1.37	.46	.21	1.70	.57	.21	1.79	1.00	.25	.12
36	1.75	- .14	.17	1.54	.12	.17	1.62	.48	.27	.22
37	2.13	.09	.16	1.88	.15	.16	1.98	.51	.19	.06
38	.97	.07	.38	.98	- .30	.38	1.03	.00	.19	.23
39	2.19	2.16	.22	.99	3.10	.22	1.04	3.91	.34	.74
40	.60	.29	.36	.70	- .30	.36	.74	- .01	.15	.38
41	.92	1.92	.24	.71	2.37	.24	.75	3.07	.32	.35
42	1.11	2.70	.31	1.01	3.26	.31	1.06	4.09	.39	.37
43	1.13	3.12	.32	1.25	2.68	.32	1.32	3.42	.34	.29
45	1.04	2.32	.27	1.23	1.85	.27	1.30	2.47	.26	.34
46	.84	.65	.15	1.00	.70	.15	1.05	1.14	.21	.14
47	.51	1.73	.21	.66	1.41	.21	.70	1.96	.24	.32
48	.80	.25	.11	1.06	.44	.11	1.12	.85	.20	.26

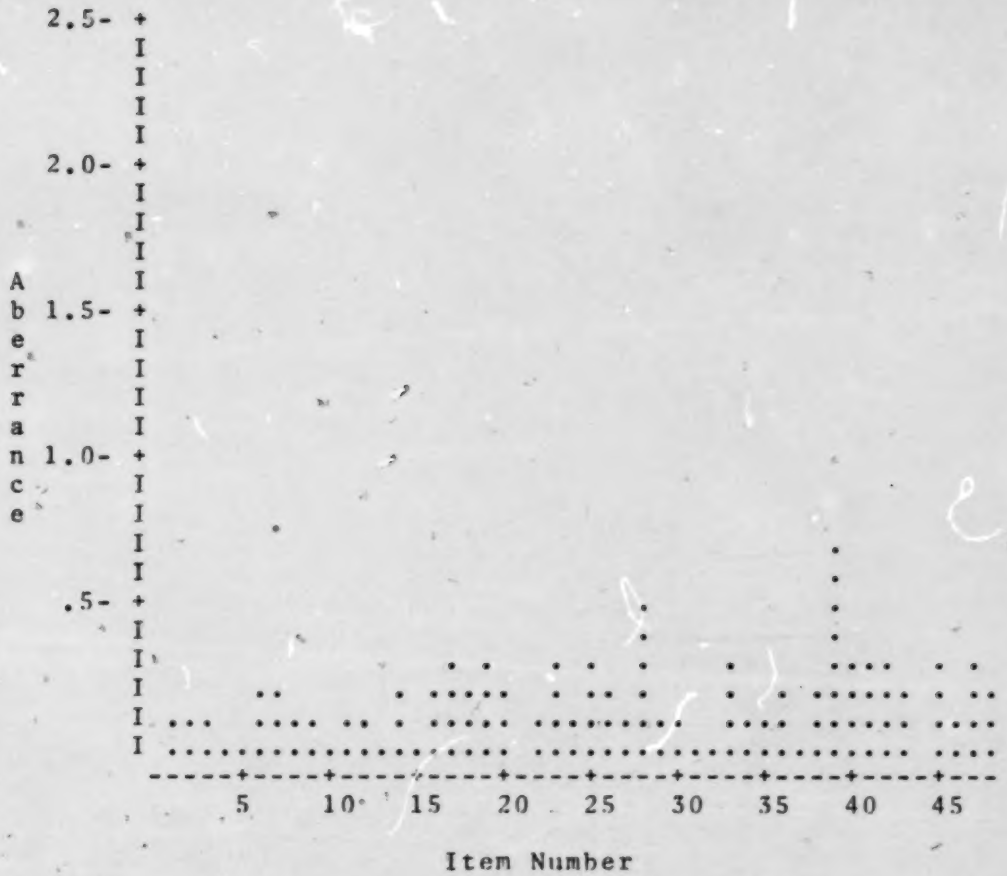

Figure 3: Plot of the degrees of item aberrance identified in the pseudo-culture group comparison.

identified degrees of aberrance are shown in Table 4. In Figure 4, the identified degrees of aberrance are plotted.

---

insert Table 3, Table 4 and Figure 4 about here

The test developer can use the identified degrees of aberrance to determine which items to consider as biased. If one wishes to screen items during test development, then a liberal definition of a very biased item, e.g. $\phi_g > .4$, could be used. (In test development it usually would be wise to liberally reject items suspected as being biased.) On the other hand, if one wishes to identify saliant characteristics of very biased items, a more conservative value, e.g. $\phi_g > .7$, could be used.

In an exploratory study in corporating these two populations and using Angoff's transformed p-value regression technique, Rudner (1977c) pooled items identified as very biased from 13 measures and found six English constructions causing undo difficulty for hearing impaired examinees. Items 4, 16, 17, 22 and 25, which had ϕ_g's >.5 and which were biased in favor of hearing examinees, fell into one of these six categories.

EYEBALLING

One could use the equated parameter values to identify biased items in a manner analagous to the procedure used by Green and Draper as illustrated in Figures 5 and 6. In Figure 5 (representing a relatively unbiased item), the icc's obtained for both groups are quite similiar. Examinees of the same true ability but different cultural affiliations have similar probabilities of responding correctly. The icc's shown in Figure 6, however, are quite different. Examinees of the same latent

Table 3

Test Statistics for the Two Diverse-Culture Groups on the SAT

	N	$\bar{x}$	s.d.	KR-20
Hearing Examinees	1603	28.9	12.44	.95
Hearing Impaired Examinees	2637	21.6	7.42	.83

Table 4

Equated and Unequated Parameter Estimates and
Degrees of Aberrance for the Diverse-Culture Group Comparison

Item #	Hearing Impaired			Hearing (Equated)			Hearing (Unequated)			Aberrance
	a	b	c	a	b	c	a	b	c	
1	1.01	- .83	.08	1.18	- .40	.08	.89	.10	.21	.40
2	1.42	- .83	.07	1.59	- .78	.07	1.20	-1.14	.32	.07
3	1.64	-1.09	.12	1.44	- .76	.12	1.09	-1.06	.29	.29
4	1.44	-1.17	.15	1.71	- .29	.15	1.29	.43	.35	.75
5	1.48	-1.14	.23	2.42	- .86	.23	1.83	-1.40	.42	.25
6	.86	- .22	.14	1.03	- .08	.14	.78	1.10	.25	.17
7	1.36	- .32	.14	1.62	- .16	.14	1.22	.86	.26	.15
8	1.27	- .59	.05	1.05	- .07	.05	.79	1.15	.31	.50
9	1.69	- .30	.02	1.09	- .22	.02	.82	.65	.17	.27
10	1.45	- .20	.13	1.69	.08	.13	1.28	1.62	.36	.24
11	1.62	.08	.12	1.22	- .30	.12	.92	.40	.19	.34
12	1.66	.05	.12	1.31	- .37	.12	.99	.19	.14	.37
13	1.72	- .62	.12	1.87	- .74	.12	1.41	-1.01	.33	.11
14	1.16	.05	.23	1.34	- .15	.23	1.01	.90	.30	.16
15	1.84	- .36	.02	1.43	- .61	.02	1.08	- .58	.24	.25
16	1.91	- .57	.07	1.40	.04	.07	1.06	1.50	.26	.57
17	1.21	- .66	.06	1.51	.15	.06	1.14	1.85	.28	.76
18	1.00	1.22	.23	1.07	.14	.23	.81	1.81	.28	.83
19	1.67	- .19	.13	.95	- .45	.13	.72	- .06	.14	.37
20	1.35	.42	.25	1.28	.20	.25	.97	2.02	.29	.16
22	.28	- .34	.00	2.25	.61	.00	1.70	3.33	.35	2.30
23	.76	- .36	.03	1.14	- .55	.03	.86	- .40	.19	.38
24	.59	- .62	.04	1.11	- .55	.04	.84	- .40	.19	.61
25	2.29	- .17	.09	.58	.33	.09	.44	2.41	.35	1.01
26	2.40	.22	.08	1.18	- .03	.08	.89	1.26	.34	.38
27	2.42	.15	.17	2.66	.18	.17	2.01	1.95	.22	.04
28	2.14	.45	.23	1.02	.38	.23	.77	2.58	.38	.32
29	2.64	.10	.01	1.46	- .08	.01	1.10	1.10	.26	.29
30	2.44	- .02	.14	1.72	.23	.14	1.30	2.10	.21	.23
31	2.12	.02	.03	1.60	- .03	.03	1.21	1.26	.27	.13
32	1.38	.08	.19	1.36	- .16	.19	1.03	.86	.23	.19
33	1.29	.29	.10	1.60	.19	.10	1.21	1.96	.32	.14
34	2.07	- .53	.05	2.16	- .37	.05	1.63	.17	.28	.15
35	2.20	- .33	.13	1.99	- .16	.13	1.50	.84	.23	.14
36	2.79	- .35	.00	2.15	- .35	.00	1.62	.25	.20	.09
37	2.15	- .35	.08	2.62	- .30	.08	1.98	.40	.16	.07
38	1.56	- .23	.11	1.38	- .38	.11	1.04	.15	.28	.14
39	1.57	.44	.08	2.69	.28	.08	2.03	2.27	.21	.23
40	.96	- .40	.04	.89	- .45	.04	.67	- .09	.17	.08
41	1.67	.62	.20	1.19	.31	.20	.90	2.37	.28	.27
42	1.50	.78	.14	1.63	.46	.14	1.23	2.84	.30	.27
43	1.32	.51	.11	1.51	.52	.11	1.14	3.04	.28	.07
45	1.72	1.07	.31	1.44	.27	.31	1.09	2.23	.21	.55
46	1.54	.12	.08	1.19	- .13	.08	.90	.96	.16	.25
47	1.63	.58	.24	.69	.09	.24	.52	1.66	.13	.60
48	2.07	.00	.04	1.19	- .23	.04	.90	.63	.12	.34

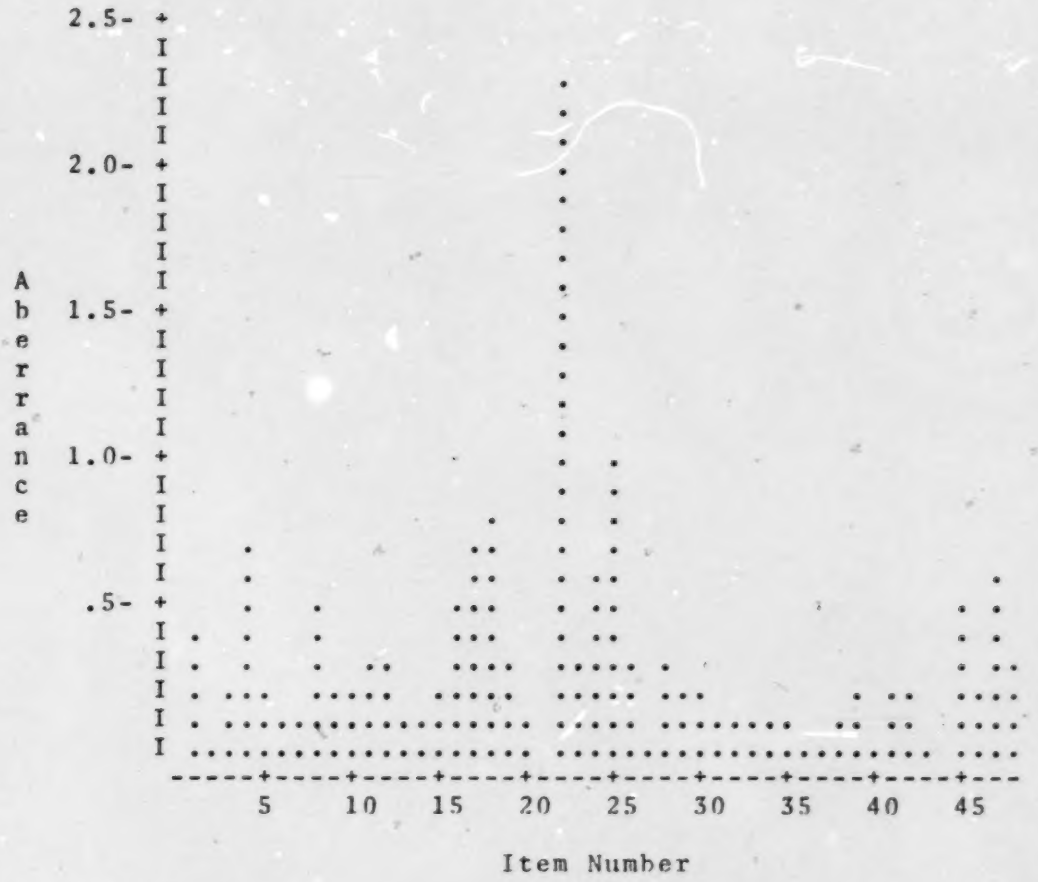

Figure 4: Plot of the degrees of item aberrance identified in the diverse-culture group comparison.

ability but from different culture groups usually have different probabilities of responding correctly. Thus, the item is relatively biased.

---
insert Figure 5 and Figure 6 about here

Eyeballing the icc's in this manner allows the researcher to get a feel for the bias. Compare Figure 7 with that of Figure 6. Both items have the approximately the same amount of bias. The item shown in Figure 6 is biased over a broad range of examinee abilities, while the item shown in Figure 7 is very biased over a narrower range. Further, eyeballing clearly illustrates which group is favored by the item. Item 18 favors hearing impaired examinees and item 17 favors hearing examinees. Thus, eyeballing offers advantages that the single numeric used to quantify bias does not.

---
insert Figure 7 about here

DISCUSSION

The reader may have noted some of the following possible objections to the icc theory approach:

1. aberrance may be indicative of things other than item bias
2. directionality of bias is not identified
3. the approach is not applicable to items with extreme p-values
4. not all items fit the latent trait theory model

The first objection items from the fact that the approach identifies items which are biased, are locally dependent, measure a trait other than that measured by the other items and/or have poor parameter estimates.

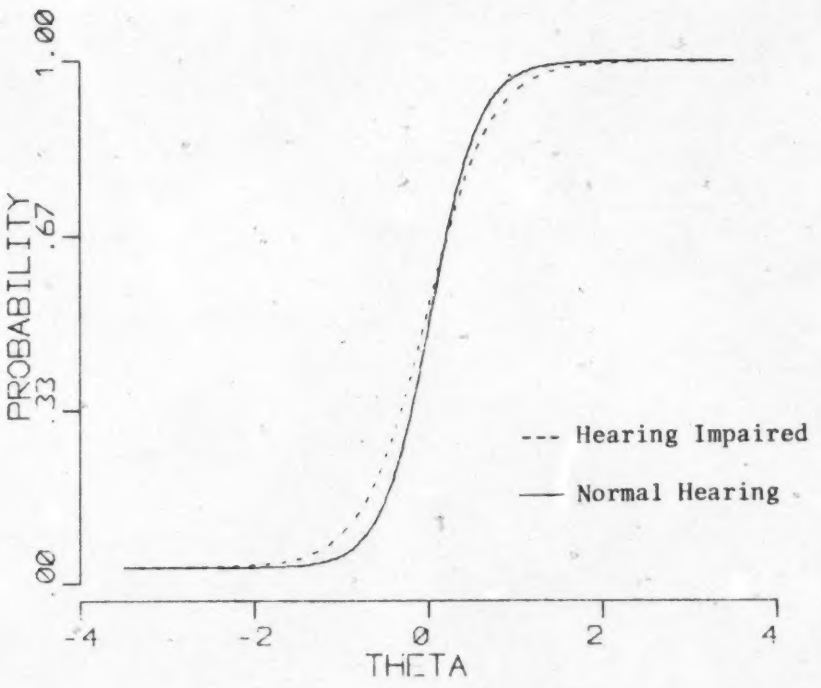

Figure 5: Estimated equated icc's for item 31

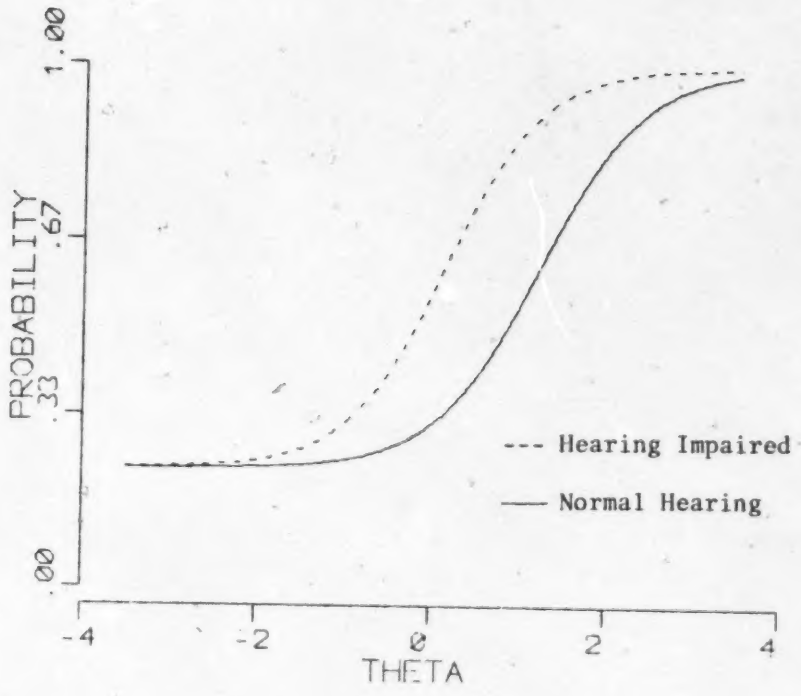

Figure 6: Estimated equated icc's for item 18

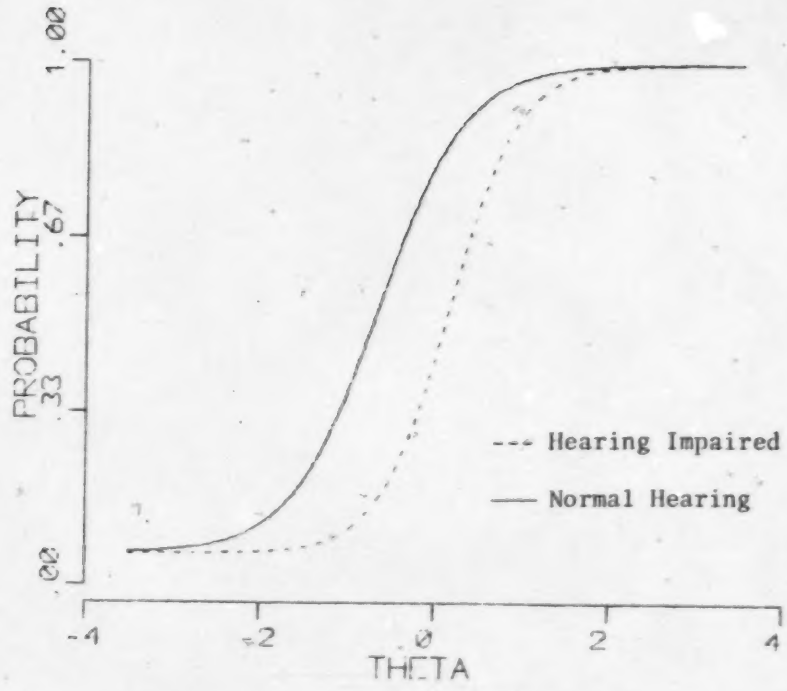

Figure 7: Estimated equated icc's for item 17

Since, in developing a measure, one would want to eliminate an item with any of the first three of these characteristics and since good parameter estimates are usually obtainable (at least when $-2 > b_g > 2$) this limitation, while existing, is felt to be relatively minor.

Even though the icc theory approach does not identify the directionality of bias, directionality can often be determined. When examinees from one culuture group consistantly have higher probabilities of responding correctly to a particular item, the item can be said to favor that group. This can be readily seen by comparing the equated b_g parameter estimates or better, by eyeballing the equated icc's. However, the reader should be aware that bias is not always directional. Consider the icc's shown in Figure 8. Low ability hearing examinees and high ability hearing impaired examinees are favored. Overall one can not say the item favors any one culture group, although a fair amount of bias is present (ϕ_g=.61). Thus, directionality is not always definable, nor should it be.

---

insert Figure 8 about here

In the pseudo-culture group two items were falsely identified as containing fair amounts of bias. Item 28 had a ϕ_g=.51 and item 39 had a ϕ_g=.74. Closer examination of these items revealed that their item difficulties were extreme. This illustrates that the icc theory approach, like many of the other approaches for biased item identification, is not always appliable to items with extreme p-values. In addition, not all items can fit the Birnbaum latent trait model. Items 21 and 44 in both comparisons could not be parameterized because of near zero or negative item - test point-biserial correlations. This indicates that ability was

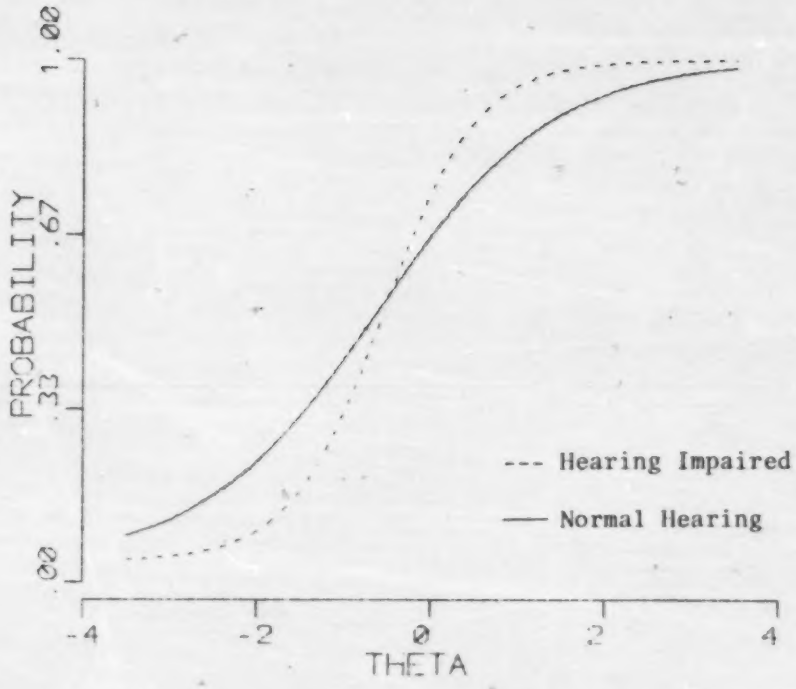

Figure 8: Estimated equated icc's for item 24

was poorly related or negatively related to the probability of a correct response. Since such items are the first to be eliminated in test development, the inability to parameterize all items does not seriously effect the utility of the approach.

Although these limitations are present, the icc theory approach appears to have several attractive properties. Most importantly, the approach utilizes a true score model- thereby lifting the tenuous assumption that observed scores are valid indicators of true ability. This was established as a criterion for an improved approach since violations of this assumption can yield spurious results.

Secondly, the approach appears to be sensitive to item bias. This contention was supported empirically in that the pseudo group comparison yielded few aberrant items and in that the actual application had identified items whose formats had previously been classified as suspect.

Third, the approach places each item on a metric to identify degree of item bias. This allows the test developer to evaluate an index of bias along with traditional indices, such as, item difficulties (e.g. p-values), discrimination indices (e.g. point biserial correlations) and dimensionality (e.g. factor loadings), to determine which items to retain for a final item pool.

Lastly, the approach is applicable to items of varying difficulty, as long as the b_g parameters are not overly extreme. Thus, the approach can be applied to most norm referenced type measures.

SUMMARY

Because it is a true score model employing item parameters which are independent of the examined sample, item characteristic curve theory offers several advantages over classical measurement theory. In this paper an approach to biased item identification using icc theory was described and applied.

The icc theory approach is attractive in that it, (1) appears to be sensitive largely to cultural variations in the trait gauged by test items, (2) does not assume total scores to be valid indicators of true ability, (3) places the identified degree of item bias on a quantified metric, and (4) is applicable to items of sufficently varying degrees of difficulty. While sensitive to some factors other than item bias, namely, local independence, item inappropriateness and poor parameter estimates, the approach may prove useful to the measurement field.

REFERENCES

Angoff, W.H., A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968, Chapts. 17-20

Green, D.R., & Draper, J.R. Exploratory studies of bias in achievement tests. Monterey: CTB/McGraw-Hill, 1972.

Jensema, C.J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished Doctoral Dissertation, University of Washington, 1972.

Jensen, A.P. An examination of culture bias in the Wonderlic Personnel Test. Arlington, VA: Eric Clearinghouse, 1973, (ERIC Document Reproduction Service No. ED 086 726.

Lord, F.M A Theory of Test Scores. Psychometric Monograph. Princeton: Educational Testing Service, 1952, (No.7).

Lord, F.M. A study of item bias using item characteristic curve theory, Proceedings of the Third Congress of Cross-Cultural Psychology, Tilburg, Holland, 1977, in press.

Lord, F.M. & Novick, M.R. Statistical Theories of Mental Test Scores. (2nd Ed.). Reading, MA: Addison-Wesley, 1974

Merz, W.R. Test fairness and test bias: A review of procedures. Paper Presented at the Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Reston, VA, May 1976.

Pine, S.M., & Weiss, D.J. Effects of item characteristics on test fairness, Research Report 76-5, Minneapolis: University of Minnesota Psychometric Methods Program, December, 1976.

Rudner, L.M. Efforts toward the development of unbiased selection and assessment instruments. Paper presented at the Third International Symposium on Educational Testing. Leyden, The Netherlands, June, 1977a.

REFERENCES

Rudner, L.M. A closer look at latent trait parameter invariance. Paper presented at the annual meeting of the New England Educational Research Organization, Manchester, NH, May 5-7, 1977b.

Rudner, L.M. Item bias with deaf and hearing examinees. Paper presented at the annual Convention of American Instructors of the Deaf, Los Angeles, June, 1977c.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April 1975.

Stokoe, W.C. The Study and Use of Sign Language. Sign Language Studies, 1976, 10, 1-36.

Urry, V.W. A monte carlo investigation of logistic mental test models. Unpublished Doctoral Dissertation, Purdue University, 1970.

Urry, V.W. Ancillary estimators for the parameters of mental test models. Paper presented at the American Psychological Association Convention, Chicago, Il, August 1975.

Wingersky, M.S., & Lord, F.M. A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses. RM 73-2. Princeton: Educational Testing Service, 1973.

Wright, B.D., Mead, R. and Draba, R., Detecting and correcting test item bias with a logistic response model (RM-22) Chicago: University of Chicago, Department of Education Statistical Laboratory, 1976.